UCRL-JRNL-203427

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Interpreting Mammalian Evolution using Fugu Genome Comparisons

I. Ovcharenko, L. Stubbs, G. G. Loots

April 7, 2004

Genomics

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Interpreting Mammalian Evolution using *Fugu* Genome Comparisons

Ivan Ovcharenko*, Lisa Stubbs and Gabriela G. Loots*

Genome Biology Department

Lawrence Livermore National Laboratory

7000 East Ave, L-441

Livermore, CA 94550

Fax: 925-422-2099

*correspondence should be addressed to:

Ivan Ovcharenko

Tel: (925) 422-5035

Email: ovcharenko1@llnl.gov

or

Gabriela G. Loots

Tel: 925-423-0923

Email: loots1@llnl.gov.

Comparative sequence analysis of the human and the pufferfish *Fugu rubripes* (fugu) genomes has revealed several novel functional coding and noncoding regions in the human genome [1, 2]. In particular, the fugu genome has been extremely valuable for identifying transcriptional regulatory elements in human loci harboring unusually high levels of evolutionary conservation to rodent genomes [3-5]. In such regions, the large evolutionary distance between human and fishes provides an additional filter through which functional noncoding elements can be detected with high efficiency.

We have evaluated the noncoding conservation profile in human/fugu genome alignments obtained from the ECR Browser (http://ecrbrowser.dcode.org/) [6] and generated by the blastz program [7]. Filtering of known and putative transcripts, pseudogenes, GenBank mRNAs, as well as proximal promoter sequences identified 2,968 human/fugu evolutionary conserved regions (ECRs) [$\geq$70% identity (%ID) over $\geq$100 basepairs (bp)] that are noncoding in nature and distantly positioned from the transcriptional start sites of adjacent genes. These ECRs are predominantly clustered in discrete areas of the human genome, flanked by or inserted into the introns of 1,026 human transcripts that together comprise only 5.6% of the 18,410 'known Gene' loci (as annotated at UCSC Genome Browser [8], hg16 freeze). This distribution suggests that human-fugu sequence comparisons will be beneficial for identifying noncoding regulatory elements for only a small percentage of human genes. Moreover, the number of genes under the control of these putative regulatory elements could be even smaller if enhancers located between two genes influence gene expression of only one of the neighboring transcripts.

It has been estimated that ~ 5% of a mammalian genome is under active selection, the majority of which will likely correspond to functional coding and noncoding sequences [9]. Human-rodent genome alignments [6] revealed 1.3 million noncoding ECRs with an average distribution of 68.8 ECRs per human gene locus, whereas the density varies according to the regional neutral substitution rates [10]. Assigning *in vivo* function to all these conserved elements is impossible with current technologies, and it is therefore critically important to identify ways to efficiently discriminate functional noncoding elements from neutrally evolving, but still highly conserved genomic DNA. This goal might be achieved if 'fingerprints' unique to functional and non-functional noncoding conserved elements can be defined. Assuming that elements conserved between human and fugu represent an incomplete yet highly enriched functional dataset, we approached this problem by studying signatures specific to human/mouse conserved noncoding elements that are also present in fishes.

We analyzed the distribution in size and percent identity of human/rodent (h/r) and human/fugu (h/f) noncoding ECRs, and compared these datasets with the h/r ECRs that are also conserved in fugu (Figure 1). Under represented regions in the mouse genome were extended by the available rat genomic sequences, in order to create a comprehensive h/r ECR dataset. The distribution in ECR length was strikingly similar between the human/mouse and human/fugu ECRs comparisons; 81% h/r and 86% h/f ECRs were shorter than 350 basepairs (bp). In sharp contrast, the majority of h/r ECRs that are conserved in fugu (90% ID) were greater than 350 bps in length. Similar striking differences were observed for the level of sequence identity. While 82% and 71% of the h/m and h/f ECRs were found to range between 70% and 77% sequence identity, 90% of h/m ECRs also conserved in fugu showed greater than 77% level of sequence identity. Therefore, our analysis suggests that a “mammalian evolutionary threshold” of ≥350bps, ≥77 % ID conservation criteria recapitulates the majority of all conserved noncoding elements identified from h/f genome comparisons, and reduces the number of h/m conserved noncoding elements 10-fold, from 1.3 millions to 128,000 ECRs, significantly simplifying the search for putative functional noncoding elements.

To correlate our findings with the conservation profiles of known regulatory elements we analyzed a 2.6 Mb region from the human *DACH* gene locus where recently seven human enhancers have been mapped [5]. Of the 1367 h/r noncoding ECRs (>100 bp/>70% ID), 34 are also present in fugu. A conservation criteria of ≥350bp/≥77% ID identifies 302 h/r ECRs and recapitulates 33/34 of the h/f conserved elements while it excludes 78% of the original h/m ECRs and maintains 100% of the experimentally validated regulatory elements [5]. Other known distant regulatory elements, including SHH and DLX1 specific developmental enhancers exceed this conservation threshold (≥350bp/≥77% ID) in h/r genomic alignments, independent of their presence in the fugu genome (Table 1) [11, 12]. We also applied these newly defined parameters on human-chicken and human-frog whole genome alignments available at the ECR Browser [6]. Over 72% of ~7500 human-frog and 55.4% of ~71,200 human-chicken noncoding ECRs that are also present in rodents obey the “mammalian evolutionary threshold” rule of conservation in the analysis of human-rodent counterpart ECRs. As we move closer in evolution within the vertebrate radiation, the significant decrease in the ratio of depicted ECRs provides a magnifying glass that allows us to visualize functional regions lacking sufficient evolutionary time to diverge in neutral regions.

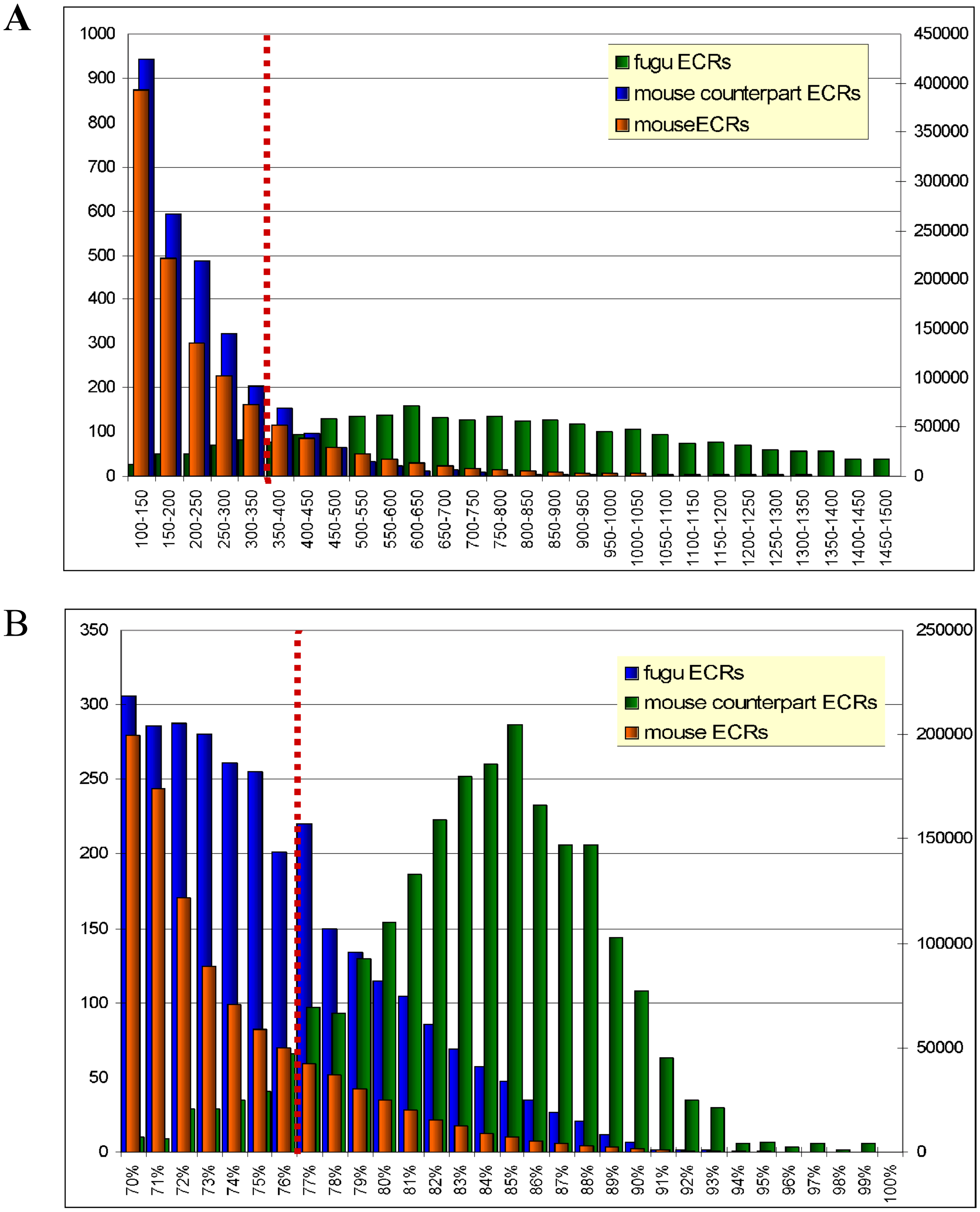
Concluding, we suggest a novel approach for analyzing human/rodent conservation profiles that is capable of reconstructing more ancestral evolutionary relationships and distinguishing functional conserved elements from the neutrally evolving genomic background. By applying a '>350bps/>77%' threshold to the analysis of human/rodent conservation profiles we were able to recapitulate the majority of human/fish conserved elements and to generate a small set of elements that have a high probability of being functional noncoding domains. Similar statistical approaches will be critical to understanding phylogenetic relationships through systematic pair-wise genomic comparisons, and has the potential to facilitate the identification of regulatory elements specific to recently evolved species such as humans and their primate relatives.

Acknowledgements: This work was supported by LLNL-LDRD 04-ERD-052. The work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory Contract No. W-7405-Eng-48.

Table 1. Experimentally characterized distant enhancer elements in the mouse.

ECR-gene	**Enhancer**	**Size (h/m)**	**H/M %ID**	**fugu cons**
Dachhund	**Nobrega MA et al. 2003**			
Dc1	negative	630	89%	y
Dc2	hindbrain	1405	89%	y
Dc3	for-, hind-brain spinal cord, retina	2458	88%	y
Dc4	Retina	1132	83%	y
Dc5	negative	730	88%	y
Dc6	midbrain, redina, drg	891	89%	y
Dc7	limb bud	1401	88%	y
Dc8	forbrain, neural tube	1023	87%	y
Dc9	Hindbrain, neural tube, genitalia	2247	82%	y
Dlx1-2	**Ghanem N et al. 2003**			
I12a	mesenchyme cells, branchyal arch	1784	84%	y
I12b	telencephalon, diencephalon	864	92%	y
Dlx5-6	**Ghanem N et al. 2003**			
mI56i	telencephalon	1477	88%	y
mI56ii	forbrain	830	88%	y
SHH	**Lettice LA et al. 2003**	1205	83%	y
Hoxc8	**Anand S et al. 2003**	583	82%	y
IL4/IL13	**Loots GG et al. 2000**	472	79%	no
FGF4	**Luster TA *et al.* 2003**	566	81%	no
pax6/nkx2.8	**Fabio Santagati et al 2003**			
cns6		500	83%	y
cns+2		1600	82%	y
pax7	**Deborah Lang et a. 2003**			
intron1		608	85%	no
ApoE	**Zheng et al. 2004**			
brain		420	75%	no

Figure 1. Genome scan of ECR length (A) and percent identity (B). Human/fugu ECRs are in blue, human/rodent ECRs are in orange, and human/rodent ECRs conserved to fugu are in green. *x*-axis, size in bp (A) and percent identity (B); *y*-axis, number of ECRs.



References:

1. Aparicio, S., et al., *Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.* Science, 2002. **297**(5585): p. 1301-10.

2. Venkatesh, B., P. Gilligan, and S. Brenner, *Fugu: a compact vertebrate reference genome.* FEBS Lett, 2000. **476**(1-2): p. 3-7.

3. Santagati, F., et al., *Identification of Cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny.* Genetics, 2003. **165**(1): p. 235-42.

4. Spitz, F., F. Gonzalez, and D. Duboule, *A global control region defines a chromosomal regulatory landscape containing the HoxD cluster.* Cell, 2003. **113**(3): p. 405-17.

5. Nobrega, M.A., et al., *Scanning human gene deserts for long-range enhancers.* Science, 2003. **302**(5644): p. 413.

6. Ovcharenko, I., Nobrega, M.A., Loots, G.G., and Stubbs, L., *ECR Browser: A Tool For Visualizing And Accessing Data From Comparisons Of Multiple Vertebrate Genomes.* Nucleic Acid Research, in press 2004.

7. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. Human-mouse alignments with BLASTZ. (2003). *Genome Res.* 13:103-7.

8. Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K.M., Pringle, T. H., Zahler, A. M., and Haussler, D. The Human Genome Browser at UCSC. (2002). *Genome Res.* 12:996-1006.

9. Margulies, E.H., et al., *Identification and characterization of multi-species conserved sequences.* Genome Res, 2003. **13**(12): p. 2507-18.

10. Hardison, R.C., et al., *Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution.* Genome Res, 2003. **13**(1): p. 13-26.

11. Lettice, L.A., et al., *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.* Hum Mol Genet, 2003. **12**(14): p. 1725-35.

12. Ghanem, N., et al., *Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters.* Genome Res, 2003. **13**(4): p. 533-43.